Outline for today

- Model of the Hessian for sum of squares loss function for two layer fully connected random net
- Random matrix theory models for the Hessian
- Introduction to Wishart and Wigner random matrices
- Stieltjes and $\mathcal{R}$ Transform for summing distributions
- Parameters where negative eigenvalues occur

Loss function for a simple two layer net

Consider a data set $X \in \mathbb{R}^{n\times m}$ of m data entries in $\mathbb{R}^n$ and associated target outputs (such as labels) $Y \in \mathbb{R}^{n_2\times m}$ (for simplicity we let $n_2 = n$). Also consider a (very) simple two layer net:

$$\begin{aligned} h_1 &= \sigma(W^{(1)}x_0) \quad \text{note, no bias, and } \sigma(\cdot) = \max(0,\cdot) \\ h_2 &= W^{(2)}h_1 \quad \text{note, no bias or nonlinear activation.} \end{aligned}$$

The output of the net is $H(x_\mu;\theta) = \hat{y}_\mu$ and we measure the value of the net through the average sum of squares:

$$\mathcal{L} = (2m)^{-1}\sum_{\mu=1}^{m}\sum_{i=1}^{n}(\hat{y}_{i,\mu} - y_{i,\mu})^2$$

and define a weighted loss accuracy as $\epsilon = n^{-1}\mathcal{L}$.

Hessian for two layer net (without activation)

Let $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ be the error in the i^{th} entry of the output for data entry indexed by μ, and $\theta = \{W^{(1)}, W^{(2)}\} \in \mathbb{R}^{2n^2}$ be the net parameters, then the hessian of the loss function has entries

$$H_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial \theta_\alpha \partial \theta_\beta} =: H_0 + H_1$$

with positive semi-definite and error dependent components:

$$[H_0]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n} \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\alpha} \frac{\partial \hat{y}_{i,\mu}}{\partial \theta_\beta} = m^{-1} [JJ^T]_{\alpha,\beta}$$

$$[H_1]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n} e_{i,\mu} \frac{\partial^2 \hat{y}_{i,\mu}}{\partial \theta_\alpha \partial \theta_\beta}.$$

Note, there are mn data entries to fit and $2n^2$ parameters in the network. Let $\phi = 2n^2/mn = 2n/m$ to measure the relative over ($\phi > 1$) or under ($\phi < 1$) parameterization.

Loss function landscape through Hessian eigenvalues

Functions, say $\mathcal{L}$, which have hessians that are:

- positive definite (all positive eigenvalues) are convex and have a single global minima and unique minimiser,
- positive semi-definite have single global minima but non-unique minimiser due to the null-space
- indefinite (positive and negative eigenvalues) are non-convex and may be a complicated landscape with multiple local minimisers.

For the simple two layer network we considered the network has Hessian $H = H_0 + H_1$ with H_0 positive semidefinite and of size independent of the error, while H_1 is indefinite with magnitude depending on the size of $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$.

Viewing the landscape through random matrix theory (Pennington et al. 17'[1])

One can interpret properties of the landscape through the Hessian by considering simplified models:

- The weights are i.i.d. random normal variable,
- The data are i.i.d. random variables,
- The residuals $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ are normal random variables, say $\mathcal{N}(0, 2\epsilon)$ with $\epsilon = n^{-1}\mathcal{L}$ (which also allows the gradient to vanish as $m \to \infty$,
- The matrices H_0 and H_1 are *freely independent* which allows us to compute the spectra of $H_0 + H_1$ from their individual spectra.

[1] http://proceedings.mlr.press/v70/pennington17a.html

Wigner and Wishart distributions

Wigner matrices, entries drawn $\mathcal{N}(0, \sigma^2)$, have eigenvalues drawn from the semi-circle law:

$$\rho_{sc}(\lambda) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2} & \text{if } |\lambda| \leq 2\sigma \\ 0 & \text{otherwise} \end{cases}$$

Wishart matrices, $X = JJ^T$ product of $J \in \mathbb{R}^{n \times p}$ drawn $\mathcal{N}(0, \sigma^2/p)$ have eigenvalues drawn from the Marchenko-Pastur distribution:

$$\rho_{MP}(\lambda) = \begin{cases} \rho(\lambda) & \text{if } \phi = n/p < 1 \\ (1 - \phi^{-1})\delta(\lambda) + \rho(\lambda) & \text{otherwise} \end{cases}$$

where $\rho(\lambda) := (2\pi\lambda\sigma\phi)^{-1}\sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)}$ for $\lambda \in [\lambda_-, \lambda_+]$ and $\lambda_\pm := \sigma(1 \pm \sqrt{\phi})^2$.

Stieltjes and $\mathcal{R}$ Transforms of probability distributions

The probability distribution of the sum of two (freely independent) random matrix distributions can be calculated using the transforms:

Stieltjes and $\mathcal{R}$ Transforms[a]

[a] https://terrytao.wordpress.com/tag/stieltjes-transform-method/

For $z \in \mathbb{C}/\mathbb{R}$ the Stieltjes Transform, $G_\rho(z)$, of a probability distribution and its inverse are given by

$$G_\rho(z) = \int_{\mathbb{R}} \frac{\rho(t)}{z-t} dt \quad \text{and} \quad \rho(\lambda) = -\pi^{-1} \lim_{\epsilon \to 0_+} Imag(G_\rho(\lambda + i\epsilon)).$$

The Stieltjes and $\mathcal{R}$ Transform of ρ are related by the solutions of $\mathcal{R}_\rho(G_\rho(z)) + 1/G_\rho(z) = z$ and has the property that if ρ_1 and ρ_2 are freely independent then $\mathcal{R}_{\rho_1+\rho_2} = \mathcal{R}_{\rho_1} + \mathcal{R}_{\rho_2}$.

Recall the Hessian for two layer net (without activation)

Let $e_{i,\mu} = \hat{y}_{i,\mu} - y_{i,\mu}$ be the error in the i^{th} entry of the output for data entry indexed by μ, and $\theta = \{W^{(1)}, W^{(2)}\} \in \mathbb{R}^{2n^2}$ be the net parameters, then the hessian of the loss function has entries

$$H_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} =: H_0 + H_1$$

with positive semi-definite and error dependent components:

$$[H_0]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n} \frac{\partial \hat{y}_{i,\mu}}{\partial\theta_\alpha} \frac{\partial \hat{y}_{i,\mu}}{\partial\theta_\beta} = m^{-1}[JJ^T]_{\alpha,\beta}$$

$$[H_1]_{\alpha,\beta} := m^{-1} \sum_{\mu=1}^{m} \sum_{i=1}^{n} e_{i,\mu} \frac{\partial^2 \hat{y}_{i,\mu}}{\partial\theta_\alpha \partial\theta_\beta}.$$

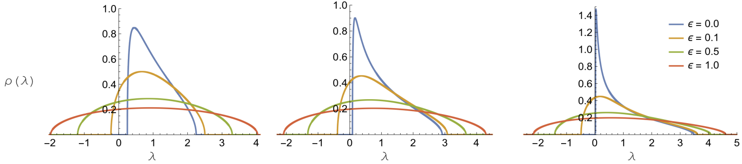
Where we assumed that H_0 and H_1 can be modelled as being drawn from Wishart and Wigner distributions respectively.

Modelling the landscape through random matrix theory (Pennington et al. 17'[2])

Using the Pennington model ($\phi = 2n/m$ and $\epsilon = n^{-1}\mathcal{L}$) we have $\rho_{H_0}(\lambda) = \rho_{MP}(\lambda; 1, \phi)$ and $\rho_{H_1}(\lambda) = \rho_{SC}(\lambda; \sqrt{2\epsilon})$.
Their $\mathcal{R}$ transforms are respectively

$$\mathcal{R}_{H_0} = \frac{1}{1 - z\phi} \quad \text{and} \quad \mathcal{R}_{H_1} = 2\epsilon z,$$

from which follows the probability distribution, $\rho_H(\lambda; \epsilon, \phi)$:



(a) $\phi = 1/4$ (b) $\phi = 1/2$ (c) $\phi = 3/4$

Figure 1. Spectral distributions of the Wishart + Wigner approximation of the Hessian for three different ratios of parameters to data points, ϕ. As the energy ϵ of the critical point increases, the spectrum becomes more semicircular and negative eigenvalues emerge.

[2] http://proceedings.mlr.press/v70/pennington17a.html

Fraction of negative eigenvalues (Pennington et al. 17'[3])

Consider the fraction of negative eigenvalues of $\rho_H(\lambda)$:

$$\alpha(\epsilon, \phi) := \int_{-\infty}^{0} \rho_H(\lambda; \epsilon, \phi) d\lambda.$$

Fraction of negative eigenvalues (without ReLU)[a]

[a] http://proceedings.mlr.press/v70/pennington17a.html

For $\rho_H(\lambda)$ modelling the Hessian of the two layer net, when α is small it is well approximated by

$$\alpha(\epsilon, \phi) \approx \alpha_0(\phi) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2}$$

where

$$\epsilon_c = \frac{1}{16}(1 - 20\phi - 8\phi^2 + (1 + 8\phi)^{3/2}).$$

[3] http://proceedings.mlr.press/v70/pennington17a.html

The two layer ReLU net (Pennington et al. 17'[4])

The introduction of the ReLU nonlinear activation changes the Hessian, roughly setting to zero half of the entries and generating a block off-diagonal structure in H_1 with $\mathcal{R}_{H1}(z) = \frac{\epsilon\phi z}{2-\epsilon\phi^2 z^2}$.
Continuing to model H_0 as Wishart (less clear an assumption):

Fraction of negative eigenvalues (with ReLU)[a]

[a] http://proceedings.mlr.press/v70/pennington17a.html

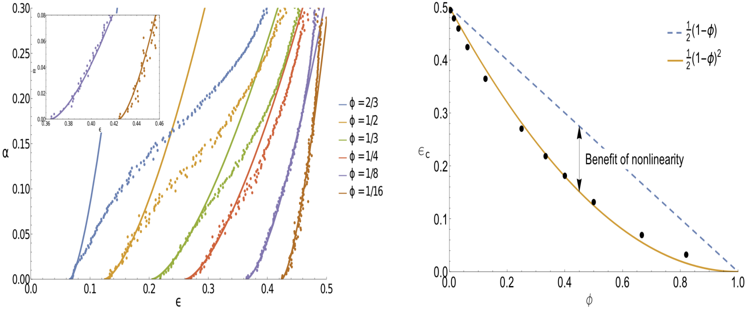
For $\rho_H(\lambda)$ modelling the Hessian of the two layer net, when α is small it is well approximated by

$$\alpha(\epsilon, \phi) \approx \tilde{\alpha}_0(\phi) \left| \frac{\epsilon - \epsilon_c}{\epsilon_c} \right|^{3/2} \quad \text{where}$$

$$\epsilon_c = \frac{\sigma^2(27 - 18\xi - \xi^2 + 8\xi^{3/2})}{32\phi(1-\phi)^3}, \quad \text{with} \quad \xi = 1 + 16\phi - 8\phi^2.$$

[4] http://proceedings.mlr.press/v70/pennington17a.html

Empirical values of ϵ_c and α (Pennington et al. 17'[5])



(a) Index of critical points versus energy

(b) Energy of minimizers versus parameters/data points

Figure 6. Empirical observations of the distribution of critical points in single-hidden-layer tanh networks with varying ratios of parameters to data points, ϕ. (a) Each point represents the mean energy of critical points with index α, averaged over ~200 training runs. Solid lines are best fit curves for small $\alpha \approx \alpha_0 |\epsilon - \epsilon_c|^{3/2}$. The good agreement (emphasized in the inset, which shows the behavior for small α) provides support for our theoretical prediction of the $3/2$ scaling. (b) The best fit value of ϵ_c from (a) versus ϕ. A surprisingly good fit is obtained with $\epsilon_c = \frac{1}{2}(1-\phi)^2$. Linear networks obey $\epsilon_c = \frac{1}{2}(1-\phi)$. The difference between the curves shows the benefit obtained from using a nonlinear activation function.

[5] http://proceedings.mlr.press/v70/pennington17a.html